ST243 Homework 3

Sicheng Zhu: 915640461

Ruoyan Yin: 916666619

Question 2

Part a

$$p(Z_i = j) = \pi_j$$

Part b

$$\begin{aligned} p(Z_i = j|x_i) &= \frac{f(x_i|Z_i = j)p(Z_i = j)}{\sum_{l=1}^{k} f(x_i|Z_i = l)p(Z_i = l)} \\ &= \frac{\frac{\pi_j}{(2\pi)^{d/2}\left|\Sigma_j\right|^{1/2}} exp\{-\frac{1}{2}\left(x_i - \mu_j\right)^T \Sigma_j^{-1}\left(x_i - \mu_j\right)\}}{\sum_{l=1}^{k} \frac{\pi_l}{(2\pi)^{d/2}|\Sigma_l|^{1/2}} exp\left\{-\frac{1}{2}(x_i - \mu_l)^T \Sigma_l^{-1}(x_i - \mu_l)\right\}} \\ &= \frac{\frac{\pi_j}{\left|\Sigma_j\right|^{1/2}} exp\{-\frac{1}{2}\left(x_i - \mu_j\right)^T \Sigma_j^{-1}\left(x_i - \mu_j\right)\}}{\sum_{l=1}^{k} \frac{\pi_l}{|\Sigma_l|^{1/2}} exp\left\{-\frac{1}{2}(x_i - \mu_l)^T \Sigma_l^{-1}(x_i - \mu_l)\right\}} \end{aligned}$$

Part c

$$F_{ij} = p_\theta\left(Z_i = j\middle|x_j\right)$$

By Jensen's Inequality,

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} \log\left[\sum_{j=1}^{k} F_{ij} \frac{p_\theta(x_i, Z_i = j)}{F_{ij}}\right] = \sum_{i=1}^{n} \log E\left[\frac{p_\theta(x_i, Z_i = j)}{q_i(Z_i)}\right] \\ &\geq \sum_{i=1}^{n} E\left[\log \frac{p_\theta(x_i, Z_i = j)}{q_i(Z_i)}\right] = \sum_{i=1}^{n}\sum_{j=1}^{k} F_{ij} \log \frac{p_\theta(x_i, Z_i = j)}{F_{ij}} \end{aligned}$$

Part d

In Jensen's Inequality, $E_\theta(\log X) = \log(E_\theta X) \iff$ X is constant to θ.

$$\begin{aligned} l(\theta') &= \sum_{i=1}^{n} \log\left[\sum_{j=1}^{k} F_{ij} \frac{p_{\theta'}(x_i, Z_i = j)}{F_{ij}}\right] \\ &= \sum_{i=1}^{n} \log\left[\sum_{j=1}^{k} p_{\theta'}(Z_i = j|x_i) \frac{p_{\theta'}(x_i, Z_i = j)}{p_{\theta'}(Z_i = j|x_i)}\right] \\ &= \sum_{i=1}^{n} \log\left[\sum_{j=1}^{k} p_{\theta'}(Z_i = j|x_i) f_{\theta'}(x_i)\right] = \sum_{i=1}^{n} \log E_{z_i}\left[f_{\theta'}(x_i)\right] \\ &= \sum_{i=1}^{n} E_{z_i}[\log f_{\theta'}(x_i)] \sum_{i=1}^{n}\sum_{j=1}^{k} p_{\theta'}(Z_i = j|x_i) \log \frac{p_{\theta'}(x_i, Z_i = j)}{p_{\theta'}(Z_i = j|x_i)} \end{aligned}$$

Part e

$$Q(\theta,\theta^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{k} F_{ij}^{(t)} \log \pi_j \, N(x_i, \mu_j, \Sigma_j)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{k} F_{ij}^{(t)} \log \pi_j + \sum_{i=1}^{n}\sum_{j=1}^{k} F_{ij}^{(t)} \, log \, N(x_i, \mu_j, \Sigma_j)$$

$$\pi^{(t+1)} = \underset{\pi:\sum \pi_j=1}{\operatorname{argmax}} \left(\sum_{i=1}^{n}\sum_{j=1}^{k} F_{ij}^{(t)} \log \pi_j \right) = \underset{\pi:\sum \pi_j=1}{\operatorname{argmax}} \sum_{i\,=\,1}^{n} \log \pi_j \left(\sum_{j=1}^{k} F_{ij}^{(t)} \right)$$

where $F_{ij}^{(t)} = p_{\theta^{(t)}}(Z_i = j|x_i) = \frac{\pi_j^{(t)} N\left(x_i, \mu_j{}^{(t)}, \Sigma_j{}^{(t)}\right)}{\sum_{j'=1}^{k} \pi_{j'}^{(t)} N\left(x_i, \mu_{j'}{}^{(t)}, \Sigma_{j'}{}^{(t)}\right)}$

Calculating $\pi_j^{(t+1)}$:

$$\phi(\pi,\lambda) = \sum_{j=1}^{k} \log \pi_j \left(\sum_{i=1}^{n} F_{ij}^{(t)} \right) + \left(\sum_{j=1}^{k} \pi_j - 1 \right)$$

$$\begin{cases} \dfrac{\partial \phi}{\partial \pi_j} = \dfrac{1}{\pi_j} \displaystyle\sum_{i=1}^{n} F_{ij}^{(t)} + \lambda = 0, \forall j = 1, \dots, k \\ \dfrac{\partial \phi}{\partial \lambda} = \displaystyle\sum_{j=1}^{k} \pi_j - 1 = 0 \end{cases} \Longrightarrow \begin{cases} \lambda = -\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{k} F_{ij}^{(t)} \\ \pi_j^{(t+1)} = \dfrac{\sum_{i=1}^{n} F_{ij}^{(t)}}{\sum_{i=1}^{n}\sum_{j=1}^{k} F_{ij}^{(t)}} = \dfrac{\sum_{i=1}^{n} F_{ij}^{(t)}}{n} \end{cases}$$

Calculating $\mu_j^{(t+1)}$ and $\Sigma_j^{(t+1)}$:

$$\left(\mu^{(t+1)}, \Sigma^{(t+1)}\right) = \sum_{i\,=\,1}^{n}\sum_{j=1}^{k} F_{ij}^{(t)} \log N\left(x_i, \mu_j, \Sigma_j\right)$$

$$= \underset{\mu,\Sigma}{\operatorname{argmax}} \sum_{i\,=\,1}^{n}\sum_{j=1}^{k} F_{ij}^{(t)} \log \frac{1}{(2\pi)^{d/2}\left|\Sigma_j\right|^{1/2}} e^{-\frac{1}{2}(x_i-\mu_j)^T \Sigma_j^{-1}(x_i-\mu_j)}$$

$$= \underset{\mu,\Sigma}{\operatorname{argmax}} \sum_{i\,=\,1}^{n}\sum_{j=1}^{k} F_{ij}^{(t)} \left(-\frac{1}{2}\left(x_i - \mu_j\right)^T \Sigma_j^{-1} \left(x_i - \mu_j\right) - \frac{1}{2}\log\left|\Sigma_j\right| \right)$$

$$= \underset{\mu,\Sigma}{\operatorname{argmax}} \sum_{i\,=\,1}^{n}\sum_{j=1}^{k} -\frac{1}{2} F_{ij}^{(t)} \left(\log\left|\Sigma_j\right| + x_i^T \Sigma_j^{-1} x_i - \mu_j^T \Sigma_j^{-1} x_i - x_i^T \Sigma_j^{-1} \mu_j \right.$$

$$\left. + \mu_j^T \Sigma_j^{-1} \mu_j \right)$$

$$\frac{\partial Q}{\partial \mu_j} = \sum_{i=1}^{n} -\frac{1}{2} F_{ij}^{(t)} \left(-\left(\Sigma_j^{-1} x_i\right)^T - x_i^T \Sigma_j^{-1} + 2\mu_j^T \Sigma_j^{-1} \right) = \sum_{i=1}^{n} F_{ij}^{(t)} \left(x_i^T \Sigma_j^{-1} - \mu_j^T \Sigma_j^{-1} \right) = 0$$

$$\sum_{i=1}^{n} F_{ij}^{(t)} \, x_i^T \Sigma_j^{-1} = \sum_{i=1}^{n} F_{ij}^{(t)} \, \mu_j^T \Sigma_j^{-1}$$

$$\Sigma_j^{-1} \sum_{i=1}^{n} F_{ij}^{(t)} \, x_i^T = \Sigma_j^{-1} \sum_{i=1}^{n} F_{ij}^{(t)} \, \mu_j^T$$

$$\Longrightarrow \mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} F_{ij}^{(t)} \, x_i}{\sum_{i=1}^{n} F_{ij}^{(t)}}$$

$$\Sigma_j^{(t+1)} = \underset{\Sigma}{\operatorname{argmax}} \sum_{i\,=\,1}^{n} -\frac{1}{2} F_{ij}^{(t)} \left(\left(x_i - \mu_j^{(t+1)} \right)^T \Sigma_j^{-1} \left(x_i - \mu_j^{(t+1)} \right) + \log\left|\Sigma_j\right| \right)$$

$$= \underset{\Sigma}{\operatorname{argmax}} \left[-\frac{1}{2} \sum_{i\,=\,1}^{n} tr\left(F_{ij}^{(t)} \Sigma_j^{-1} \left(x_i - \mu_j^{(t+1)} \right) \left(x_i - \mu_j^{(t+1)} \right)^T \right) - \frac{1}{2} \log\left|\Sigma_j\right| \sum_{i=1}^{n} F_{ij}^{(t)} \right]$$

$$= \underset{\Sigma}{\operatorname{argmax}} \left[-\frac{1}{2} tr\left(\Sigma_j^{-1} \sum_{i=1}^{n} F_{ij}^{(t)} \left(x_i - \mu_j^{(t+1)} \right) \left(x_i - \mu_j^{(t+1)} \right)^T \right) - \frac{1}{2} \log\left|\Sigma_j\right| \sum_{i=1}^{n} F_{ij}^{(t)} \right]$$

$$\frac{\partial Q}{\partial \Sigma_j} = -\frac{1}{2} \left[-\left(\Sigma_{\mathrm{j}}^{-1}\right)^{\mathrm{T}} \sum_{i=1}^{n} F_{ij}^{(t)} \left(x_i - \mu_j^{(t+1)} \right) \left(x_i - \mu_j^{(t+1)} \right)^T \left(\Sigma_j^{-1} \right)^T + \Sigma_j^{-1} \sum_{i=1}^{n} F_{ij}^{(t)} \right] = 0$$

As Σ_j is diagonal matrix, Σ_{j}^{-1} is also diagonal matrix, $\left(\Sigma_{\mathrm{j}}^{-1}\right)^{\mathrm{T}} = \Sigma_{\mathrm{j}}^{-1}$

$$\Sigma_j^{-1} \sum_{i=1}^{n} F_{ij}^{(t)} \left(x_i - \mu_j^{(t+1)} \right) \left(x_i - \mu_j^{(t+1)} \right)^T \Sigma_j^{-1} = \sum_{i=1}^{n} F_{ij}^{(t)} \Sigma_j^{-1}$$

$$\Sigma_j \Sigma_j^{-1} \sum_{i=1}^{n} F_{ij}^{(t)} \left(x_i - \mu_j^{(t+1)} \right) \left(x_i - \mu_j^{(t+1)} \right)^T \Sigma_j^{-1} \Sigma_j = \Sigma_j \sum_{i=1}^{n} F_{ij}^{(t)} \Sigma_j^{-1} \, \Sigma_j$$

$$\sum_{i=1}^{n} F_{ij}^{(t)} \left(x_i - \mu_j^{(t+1)} \right) \left(x_i - \mu_j^{(t+1)} \right)^T = \Sigma_j \sum_{i=1}^{n} F_{ij}^{(t)}$$

$$\Longrightarrow \Sigma_j^{(t+1)} = \frac{\sum_{i=1}^{n} F_{ij}^{(t)} \left(x_i - \mu_j^{(t+1)} \right) \left(x_i - \mu_j^{(t+1)} \right)^T}{\sum_{i=1}^{n} F_{ij}^{(t)}}$$

$$\begin{cases} \pi_j^{(t+1)} = \dfrac{\sum_{i=1}^{n} F_{ij}^{(t)}}{\mathrm{n}} \\ \mu_j^{(t+1)} = \dfrac{\sum_{i=1}^{n} F_{ij}^{(t)} x_i}{\sum_{i=1}^{n} F_{ij}^{(t)}} \\ \Sigma_j^{(t+1)} = \dfrac{\sum_{i=1}^{n} F_{ij}^{(t)} \left(x_i - \mu_j^{(t+1)}\right)\left(x_i - \mu_j^{(t+1)}\right)^T}{\sum_{i=1}^{n} F_{ij}^{(t)}} \end{cases}$$

In a mixture of spherical Gaussians model, $\Sigma_j = {\sigma_j}^2 I_d, |\Sigma_j| = {\sigma_j}^{2d}, {\Sigma_j}^{-1} = \frac{1}{{\sigma_j}^2} I_d$

$$N\left(x_i, \mu_j, \Sigma_j\right) = \frac{1}{(2\pi)^{d/2}\left|\Sigma_j\right|^{1/2}} e^{-\frac{1}{2}\left(x_i-\mu_j\right)^T \Sigma_j^{-1}\left(x_i-\mu_j\right)}$$

$$= \frac{1}{(2\pi)^{d/2}{\sigma_j}^d} e^{-\frac{1}{2}\left(x_i-\mu_j\right)^T \frac{1}{{\sigma_j}^2} I_d \left(x_i-\mu_j\right)} = \frac{1}{(2\pi)^{d/2}{\sigma_j}^d} e^{-\frac{1}{2{\sigma_j}^2}\left(x_i-\mu_j\right)^T \left(x_i-\mu_j\right)}$$

So,

$$F_{ij}^{(t)} = \frac{\pi_j^{(t)} N\left(x_i, {\mu_j}^{(t)}, {\Sigma_j}^{(t)}\right)}{\sum_{j'=1}^{k} \pi_{j'}^{(t)} N\left(x_i, {\mu_{j'}}^{(t)}, {\Sigma_{j'}}^{(t)}\right)}$$

$$= \frac{\pi_j^{(t)} \frac{1}{(2\pi)^{d/2}{\sigma_j}^{(t)^d}} exp\{-\frac{1}{2{\sigma_j}^{(t)^2}}\left(x_i - {\mu_j}^{(t)}\right)^T \left(x_i - {\mu_j}^{(t)}\right)\}}{\sum_{j'=1}^{k} \pi_{j'}^{(t)} \frac{1}{(2\pi)^{d/2}{\sigma_{j'}}^{(t)^d}} exp\{-\frac{1}{2{\sigma_{j'}}^{(t)^2}}\left(x_i - {\mu_{j'}}^{(t)}\right)^T \left(x_i - {\mu_{j'}}^{(t)}\right)\}}$$

$$= \frac{\pi_j^{(t)} \frac{1}{{\sigma_j}^{(t)^d}} exp\{-\frac{1}{2{\sigma_j}^{(t)^2}}\left(x_i - {\mu_j}^{(t)}\right)^T \left(x_i - {\mu_j}^{(t)}\right)\}}{\sum_{j'=1}^{k} \pi_{j'}^{(t)} \frac{1}{{\sigma_{j'}}^{(t)^d}} exp\{-\frac{1}{2{\sigma_{j'}}^{(t)^2}}\left(x_i - {\mu_{j'}}^{(t)}\right)^T \left(x_i - {\mu_{j'}}^{(t)}\right)\}}$$

$$\sigma_j^{(t+1)} = \underset{\sigma}{\operatorname{argmax}} \sum_{i=1}^{n} -\frac{1}{2} F_{ij}^{(t)} \left(\frac{1}{{\sigma_j}^2}\left(x_i - \mu_j^{(t+1)}\right)^T \left(x_i - \mu_j^{(t+1)}\right) + \log {\sigma_j}^{2d}\right)$$

$$= \underset{\sigma}{\operatorname{argmax}} \sum_{i=1}^{n} -\frac{1}{2} F_{ij}^{(t)} \frac{1}{{\sigma_j}^2}\left(x_i - \mu_j^{(t+1)}\right)^T \left(x_i - \mu_j^{(t+1)}\right) - d\, F_{ij}^{(t)} \log \sigma_j$$

$$\frac{\partial \mathrm{Q}}{\partial \sigma_{\mathrm{j}}} = \frac{1}{{\sigma_j}^3} \sum_{i=1}^{n} F_{ij}^{(t)} \left(x_i - \mu_j^{(t+1)}\right)^T \left(x_i - \mu_j^{(t+1)}\right) - \frac{d}{\sigma_j} \sum_{i=1}^{n} F_{ij}^{(t)} = 0$$

$$\begin{cases} \pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} F_{ij}^{(t)}}{\mathrm{n}} \\ \mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} F_{ij}^{(t)} x_i}{\sum_{i=1}^{n} F_{ij}^{(t)}} \\ \sigma_j^{2(t+1)} = \frac{\sum_{i=1}^{n} F_{ij}^{(t)} \left(x_i - \mu_j^{(t+1)}\right)^T \left(x_i - \mu_j^{(t+1)}\right)}{d \sum_{i=1}^{n} F_{ij}^{(t)}} \end{cases},$$

where $F_{ij}^{(t)} = \frac{\pi_j^{(t)} \frac{1}{{\sigma_j}^{(t)^d}} exp\{-\frac{1}{2{\sigma_j}^{(t)^2}} (x_i - {\mu_j}^{(t)})^T (x_i - {\mu_j}^{(t)})\}}{\sum_{j'=1}^{k} \pi_{j'}^{(t)} \frac{1}{{\sigma_{j'}}^{(t)^d}} exp\{-\frac{1}{2{\sigma_{j'}}^{(t)^2}} \left(x_i - {\mu_{j'}}^{(t)}\right)^T \left(x_i - {\mu_{j'}}^{(t)}\right)\}}$

Part f

In a mixture of diagonal Gaussians model, $\Sigma_j = diag\left({\sigma_{j1}}^2, {\sigma_{j2}}^2, \dots, {\sigma_{jd}}^2\right)$,

$|\Sigma_j| = \prod_{l=1}^{d} {\sigma_{jl}}^2, {\Sigma_j}^{-1} = diag\left(\frac{1}{{\sigma_{j1}}^2}, \frac{1}{{\sigma_{j2}}^2}, \dots, \frac{1}{{\sigma_{jd}}^2}\right)$

Let ${\sigma_j}^2 = [{\sigma_{j1}}^2, {\sigma_{j2}}^2, \dots, {\sigma_{jd}}^2]$, and $\frac{1}{{\sigma_j}^2} = [\frac{1}{{\sigma_{j1}}^2}, \frac{1}{{\sigma_{j2}}^2}, \dots, \frac{1}{{\sigma_{jd}}^2}]$

$$N\left(x_i, \mu_j, \Sigma_j\right) = \frac{1}{(2\pi)^{d/2} \left|\Sigma_j\right|^{1/2}} e^{-\frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j)}$$

$$= \frac{1}{(2\pi)^{d/2} \prod_{l=1}^{k} \sigma_{jl}} e^{-\frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j)}$$

$$= \frac{1}{(2\pi)^{d/2} \prod_{l=1}^{k} \sigma_{jl}} e^{-\frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j)}$$

$$= \frac{1}{(2\pi)^{d/2} \prod_{l=1}^{d} \sigma_{jl}} exp\{\sum_{l=1}^{d} -\frac{1}{2{\sigma_{jl}}^2} \left(x_{il} - \mu_{jl}\right)^T \left(x_{il} - \mu_{jl}\right)\}$$

So,

$$F_{ij}^{(t)} = \frac{\pi_j^{(t)} N\left(x_i, {\mu_j}^{(t)}, {\Sigma_j}^{(t)}\right)}{\sum_{j'=1}^{k} \pi_{j'}^{(t)} N\left(x_i, {\mu_{j'}}^{(t)}, {\Sigma_{j'}}^{(t)}\right)}$$

$$= \frac{\pi_j^{(t)} \frac{1}{(2\pi)^{d/2} \prod_{l=1}^{k} {\sigma_{jl}}^{(t)}} exp\{\sum_{l=1}^{d} -\frac{1}{2{\sigma_{jl}}^{(t)^2}} \left(x_{il} - \mu_{jl}\right)^2\}}{\sum_{j'=1}^{k} \pi_{j'}^{(t)} \frac{1}{(2\pi)^{d/2} \prod_{l=1}^{k} {\sigma_{j'l}}^{(t)}} exp\{\sum_{l=1}^{d} -\frac{1}{2{\sigma_{j'l}}^{(t)^2}} \left(x_{il} - \mu_{j'l}\right)^2\}}$$

$$= \frac{\pi_j^{(t)} \frac{1}{\prod_{l=1}^{k} \sigma_{jl}{}^{(t)}} exp\{\sum_{l=1}^{d} -\frac{1}{2\sigma_{jl}{}^{(t)2}} \left(x_{il} - \mu_{jl}\right)^2\}}{\sum_{j'=1}^{k} \pi_{j'}^{(t)} \frac{1}{\prod_{l=1}^{k} \sigma_{j'l}{}^{(t)}} exp\{\sum_{l=1}^{d} -\frac{1}{2\sigma_{j'l}{}^{(t)2}} \left(x_{il} - \mu_{j'l}\right)^2\}}$$

$$\sigma_{jl}^{(t+1)} = \underset{\sigma_{jl}}{\operatorname{argmax}} \sum_{i=1}^{n} -\frac{1}{2\sigma_{jl}{}^2} F_{ij}^{(t)} \left(\left(x_{il} - \mu_{jl}^{(t+1)}\right)^2 + 2\log\sigma_{jl}\right)$$

$$= \underset{\sigma_{jl}}{\operatorname{argmax}}[-\frac{1}{2}\sum_{i=1}^{n} \frac{1}{\sigma_{jl}{}^2} F_{ij}^{(t)} \left(x_{il} - \mu_{jl}^{(t+1)}\right)^2 - \frac{1}{2}\log\sigma_{jl} \sum_{i=1}^{n} F_{ij}^{(t)}]$$

$$\frac{\partial Q}{\partial \sigma_{jl}} = \frac{1}{\sigma_{jl}{}^3} \sum_{i=1}^{n} F_{ij}^{(t)} \left(x_{il} - \mu_{jl}^{(t+1)}\right)^2 - \frac{1}{2\sigma_{jl}} \sum_{i=1}^{n} F_{ij}^{(t)} = 0$$

$$\begin{cases} \pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} F_{ij}^{(t)}}{n} \\ \mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} F_{ij}^{(t)} x_i}{\sum_{i=1}^{n} F_{ij}^{(t)}} \\ \sigma_{jl}^{2(t+1)} = \frac{2\sum_{i=1}^{n} F_{ij}^{(t)} \left(x_{il} - \mu_{jl}^{(t+1)}\right)^2}{\sum_{i=1}^{n} F_{ij}^{(t)}}, l = 1, \ldots, d \end{cases},$$

where $F_{ij}^{(t)} = \frac{\pi_j^{(t)} \frac{1}{\prod_{l=1}^{k} \sigma_{jl}{}^{(t)}} exp\{\sum_{l=1}^{d} -\frac{1}{2\sigma_{jl}{}^{(t)2}} \left(x_{il} - \mu_{jl}\right)^2\}}{\sum_{j'=1}^{k} \pi_{j'}^{(t)} \frac{1}{\prod_{l=1}^{k} \sigma_{j'l}{}^{(t)}} exp\{\sum_{l=1}^{d} -\frac{1}{2\sigma_{j'l}{}^{(t)2}} \left(x_{il} - \mu_{j'l}\right)^2\}}$

Question 3

(i)

The EM algorithm for mixture of spherical Gaussians is implemented. And because each x_i is an 1x196 row vector, the value of $f(x_i, \mu_j, \Sigma_j)$ could be extremely small. In order to get the precise value of $F_{ij}^{(t)}$, we apply the log-sum-exp trick as below:

$$\log F_{ij}^{(t)} = \log\left(\pi_j^{(t)} N\left(x_i, \mu_j{}^{(t)}, \Sigma_j{}^{(t)}\right)\right) - \log\left(\sum_{j'=1}^{k} \pi_{j'}^{(t)} N\left(x_i, \mu_{j'}{}^{(t)}, \Sigma_{j'}{}^{(t)}\right)\right)$$

$$= \log\left(\pi_j^{(t)} N\left(x_i, \mu_j{}^{(t)}, \Sigma_j{}^{(t)}\right)\right) - \log\left(\sum_{j'=1}^{k} exp\{\log(\pi_{j'}^{(t)} N\left(x_i, \mu_{j'}{}^{(t)}, \Sigma_{j'}{}^{(t)}\right))\}\right)$$

$$= \log\left(\pi_j^{(t)} N\left(x_i, \mu_j{}^{(t)}, \Sigma_j{}^{(t)}\right)\right) - [a + \log \sum_{j'=1}^{k} exp\{\log(\pi_{j'}^{(t)} N\left(x_i, \mu_{j'}{}^{(t)}, \Sigma_{j'}{}^{(t)}\right)) - a\}]$$

where $a = \max_j \log\left(\pi_j^{(t)} N\left(x_i, \mu_j{}^{(t)}, \Sigma_j{}^{(t)}\right)\right)$,

$$\log\left(\pi_j^{(t)} N\left(x_i, \mu_j{}^{(t)}, \Sigma_j{}^{(t)}\right)\right) = \log\pi_j^{(t)} - d\log\sigma_j{}^{(t)} - \frac{1}{2\sigma_j{}^{(t)^2}}\left(x_i - \mu_j{}^{(t)}\right)^T\left(x_i - \mu_j{}^{(t)}\right).$$

Then, within each iteration, we calculate the matrix F, and compute $\pi_j^{(t+1)}, \mu_j{}^{(t+1)}, \sigma_j{}^{(t+1)}$ using the result of **part e** of Question 2.

As we need the fractional change of the log-likelihood between each iteration to decide terminating or not, for the first iteration, in order to let $\frac{log-likelihood_{(t)} - log-likelihood_{(t-1)}}{log-likelihood_{(t-1)}}$ work well for t = 1, we set $log - likelihood_{(0)} = 1$, to make sure the denominator is not 0.

We tried 3 initializations:

	I	II	III
$\pi^{(0)}$	[0.2 0.2 0.2 0.2 0.2]	[0.1 0.4 0.4 0.02 0.08]	[0.8 0.05 0.05 0.05 0.05]
$\mu_{jl}{}^{(0)}$	~ N(100, 10)	~ N(50, 5)	~ N(0, 1)
$\sigma_j{}^{(0)}$	100	~ N(50, 10)	~ N(200, 10)

The iteration times and log-likelihoods results are as below:

	I	II	III
Iteration times	12	19	11
log-likelihood	-31827008	-31795742	-31827819

The second iteration got the best result in terms of maximum log-likelihood. The third iteration got the best result in terms of speed of convergence.
We identify the digits of clusters by the following method:
(a) Assume the digit of the cluster to be i;
(b) Count the number of correctly classified observations t;
(c) Compare t of different i's, and assign the cluster to be digit i with largest t.
We plot each μ_j and the digits showing in the below images are exactly the same as what we identity from the above methods :

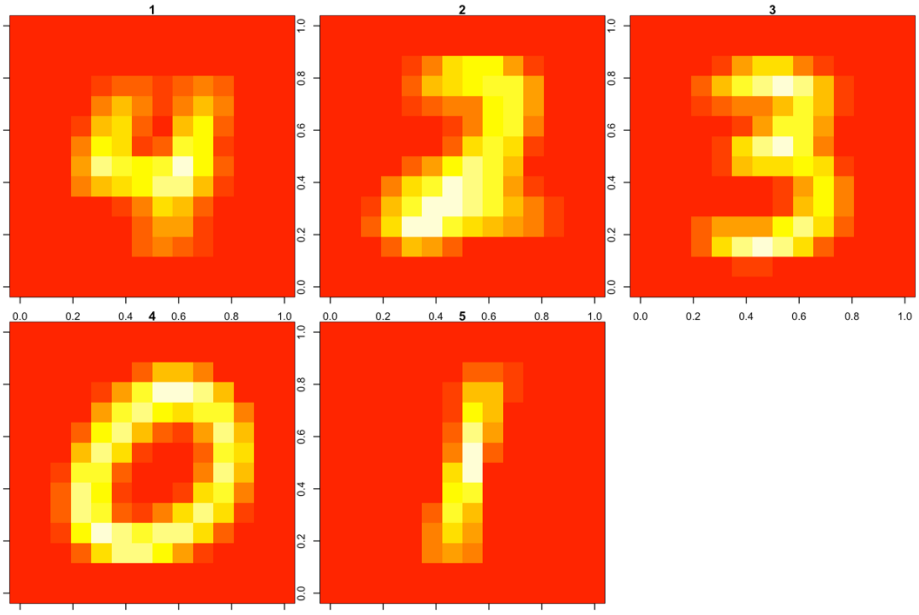

Figure 1 Initialization I

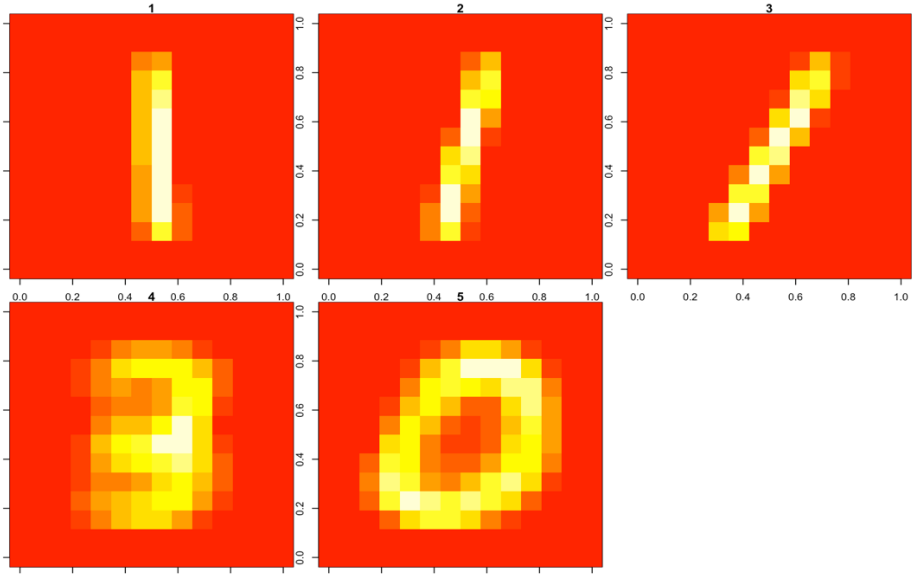

Figure 2 Initialization II

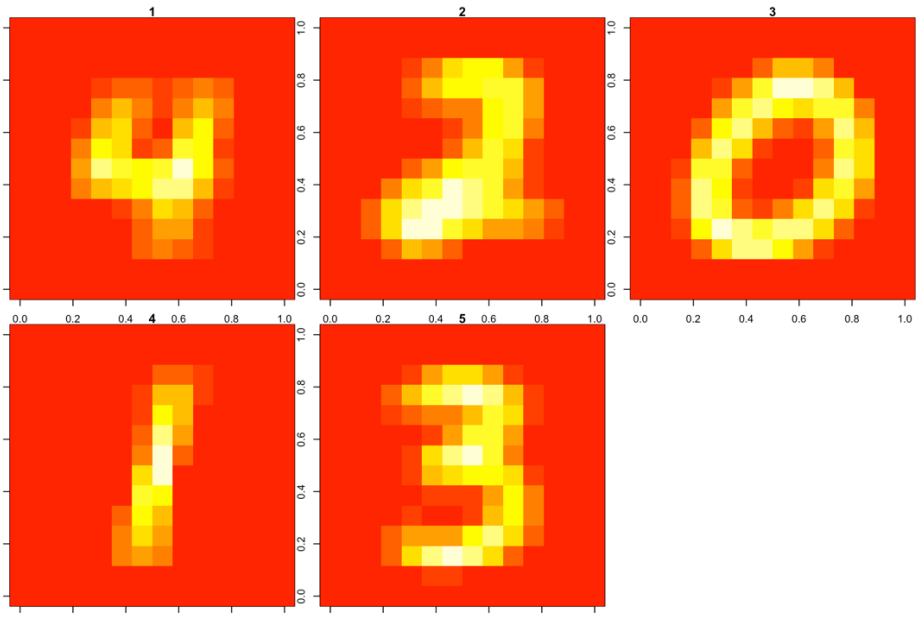

Figure 3 Initialization III

For initialization I and III, the clusters are successful, as five clusters refer to five digits. But for initialization II, 3 clusters are digit 1, and it misses digits 2 and 4. Although the log-likelihood from initialization II is minimal among all the three, the result of it is bad, and the errors should be very large such that we do not test over this model.

Using EM algorithms from the three initializations to fit the testing data, the errors are as below:

	I	III
Errors	0.1152	0.1136

The clusters coming from initialization III has the best result in terms of prediction error. Still, the error is significantly large, so t a mixture models of Gaussian distributions may not be suitable for modeling the MNIST data.

(ii)

Then we applied EM algorithm for diagonal Gaussians. Still, we used log-exp trick to avoid the problem that the conditional probability and the log-likelihood can be underflow. Just like what we did in part (i), we still set the proportional change in log-likelihood as criterion to make the decision if the classification is good enough.

Since the outcome can be highly dependent on the initialized parameters, we tried three different combinations.

	I	II	III
π	(0.2 0.25 0.25 0.15 0.15)	(0.15 0.20 0.25 0.20 0.20)	(0.3 0.2 0.1 0.20 0.20)
μ	~N(80, 2)	~N(100, 10)	~N(60,5)
σ	uniform(0.9,1.1)	uniform(0.5, 1.3)	uniform(0.8, 1.5)

Chart below shows the iteration numbers taken by different combinations and the log-likelihood based on those initializations.

	I	II	III
Iteration times	7	11	11
log-likelihood	-21485205	-19067068	--19658301

Even the second combination get the best result with respect to the log-likelihood, it didn't successfully depart the data into 5 groups. And the third one got a better likelihood even it takes a few steps than the first.

Using EM algorithms from the three initializations to fit the testing data, the errors are as below:

	I	III
Errors	0.5606149	0.5197509

The clusters coming from initialization III has a relative better result in terms of prediction error compared to I. However, the error is significantly large, so a mixture models of Gaussian distributions may not be suitable for modeling the MNIST data.